ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9415096

END-TO-END LYRICS RECOGNITION WITH VOICE TO SINGING STYLE TRANSFER

Sakya Basak[$], *Shrutina Agarwal*[$], *Sriram Ganapathy*[$], *Naoya Takahashi*[*]

[$]Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore, India,
[*]Sony Corporation, Tokyo, Japan.

ABSTRACT

Automatic transcription of monophonic/polyphonic music is a challenging task due to the lack of availability of large amounts of transcribed data. In this paper, we propose a data augmentation method that converts natural speech to singing voice based on vocoder based speech synthesizer. This approach, called voice to singing (V2S), performs the voice style conversion by modulating the F0 contour of the natural speech with that of a singing voice. The V2S model based style transfer can generate good quality singing voice thereby enabling the conversion of large corpora of natural speech to singing voice that is useful in building an E2E lyrics transcription system. In our experiments on monophonic singing voice data, the V2S style transfer provides a significant gain (relative improvements of 21 %) for the E2E lyrics transcription system. We also discuss additional components like transfer learning and lyrics based language modeling to improve the performance of the lyrics transcription system.

Index Terms— Voice-to-singing style transfer, Lyrics Transcription, End-to-end modeling.

1. INTRODUCTION

In music, lyrics constitutes the textual component of singing voice. It forms an important constituent of the music signal that contributes to the emotional perception of the song [1] and also aids in foreign language learning [2]. In music information extraction, two problems of interest are the automatic alignment of lyrics and the automatic transcription of singing voice. The alignment problem is the task of finding the timing of the word boundaries for the given lyrics with respect to the polyphonic audio [3], while transcription is the task of recognizing the lyrics [4]. Several applications such as generating karaoke, music subtitling [5], query-by-singing [6], keyword spotting, and automatic indexing of music according to transcribed keywords [7] rely on accurate alignment and transcription of music.

The key challenges in the automatic recognition of lyrics is the unique style of singing voice, high variation of fundamental frequency and pronunciation, lack of large amounts of transcribed singing data and background scores. The earlier studies used a phoneme recognition approach [8]. Mesaros et al. [4] adopted an automatic speech recognition (ASR) based approach for phoneme and word recognition of singing vocals in monophonic and polyphonic music. In dealing with polyphonic music, one of the common approaches is to apply a voice source separation module [9]. In a recent work, Gupta et. al. [3] attempted an adaptation of the models trained from solo music to polyphonic music for lyrics alignment.

In this paper, we attempt to perform automatic recognition of lyrics by utilizing the large amount of resources available for speech recognition. The conventional modular approach to ASR consists of several modules like acoustic model, lexical model and language model [10]. The end-to-end (E2E) ASR is a simplified approach to overcome the limitations of the conventional approach by using a single neural network to perform an entirely data driven learning. It consists of a single deep neural network (DNN) model which is directly trained on words, sub-words or character targets, thereby eliminating the need for the hand-crafted pronunciation dictionary. The earliest approach to E2E ASR used the connectionist temporal cost (CTC) function to optimize the recurrent neural network model (RNN) [11]. The attention based models proposed recently do not make any conditional independence assumption and they attempt to learn an implicit language model using an encoder-decoder attention framework [12]. In order to combine the best of both worlds Watanabe et. al. proposed a hybrid CTC-attention model [13]. In the last year, the performance of E2E models have been improved with the use of Transformer based architectures [14]. The performance of the E2E ASR models can be further improved by using data augmentation techniques on the input features using methods like time warping, frequency masking, and time masking [15]. However, E2E ASR tends to be data demanding in training, which makes it difficult to adopt this framework for lyrics transcription tasks as there is a considerable lack of large supervised datasets.

In this paper, we propose a novel approach to data augmentation for end-to-end recognition of lyrics in singing voice. The proposed approach, termed as voice-to-singing (V2S), converts natural speech to singing voice using a vocoder based speech synthesizer [16]. The V2S model uses the pitch contours from singing voice recordings along with the spectral envelope of the natural speech to perform voice to singing conversion. The proposed V2S approach can generate large amounts of "singing" voice for use in E2E model training. In addition, we investigate a transfer learning approach to leverage a large "singing" speech trained model. The use of source-separated singing voice data from polyphonic music, which is relatively easy to obtain compared to monophonic singing voice data, is also explored. We also develop a language model (LM) suitable for lyrics transcription by mining large text corpus of lyrics. The experiments are performed on the Dali corpus [17] and a proprietary music dataset provided by Sony. In these experiments, we show that the proposed V2S approach provides significant performance gains over baseline systems trained purely on natural speech.

The key contributions from this work are,

i We propose V2S approach for data augmentation to train the E2E lyric transcription model.

ii We investigate a transfer learning approach to leverage a large speech corpus and source-separated singing voice data from polyphonic music for E2E system.

iii We develop a language model for lyrics transcription by mining

This work was funded by Sony Corporation, Tokyo.

large text corpus of lyrics.

iv Experimental results on polyphonic and monophonic lyric transcription shows that the proposed V2S data augmentation, transfer learning using speech, source-separated singing voice data, speed perturbation and lyrics LM significantly improve the word error rate over the baseline system trained on natural speech.

To the best of our knowledge, this paper constitutes one of the earliest efforts in developing E2E systems for automatic lyrics transcription of singing voice.

2. RELATED PRIOR WORK

Recently, Gupta et al. [18] explored singing voice alignment on a cappella singing vocals using state of the art Google ASR model. The authors segmented the audio in 10s segments, and the ASR model is used to get the time aligned transcription for the audio segment. The transcriptions are later refined using published lyrics and are used in training a conventional ASR. An iterative process is thereby ensued to refine the alignments. For polyphonic music, Sharma et al. [19] developed a lyrics alignment system, where the ASR models were adapted to singing voices using speaker adaptive training on a small dataset of solo singing voices. In addition to this, the authors use source separated vocals from polyphonic music using convolutional neural network (CNN) based U-Net models. In a recent work on E2E models, D.Stoller et al [20] investigated Wave-U-Net to predict character probabilities using raw audio. The short segment audio snippets (10s) are processed with the end-to-end model which uses CNN with CTC training. There has also been efforts in exploring additional features like voicing, energy, auditory, spectral, and chroma during training of the model for alignment tasks [21].

The background music may affect the intelligibility of the lyrics. Gupta et al. [22] explored the significance of background music by performing genre based learning. Here, the music is divided into three different genres (hiphop, metal, and pop) and for each genre different phoneme and silence (non-vocal) models are trained using a standard HMM-GMM architecture. The authors report improvements using the genre specific modeling.

With the control of several acoustic features, Saitou et. al [23] proposed voice conversion approach to transform speech to singing style. An encoder-decoder approach to model-based learning of speech to singing voice conversion was explored in Parekh [24]. However many of the past methods are not scalable to large volumes of speech data needed for acoustic model training in a E2E lyrics transcription system.

3. PROPOSED METHODS

In this section, we describe our approach on the V2S data augmentation, the transfer learning using source separated singing voice, and the language model development for lyrics.

3.1. Voice to Singing (V2S)

The V2S converts natural speech to singing voice using a vocoder based speech synthesizer. We use the WORLD synthesizer [16] since it provide high-quality voice with low computational complexity. We first describe the WORLD model, followed by a description of the proposed V2S.

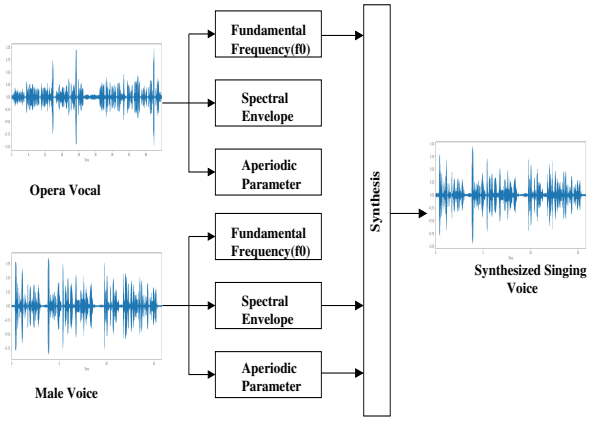


Fig. 1: WORLD vocoder used in the proposed V2S system

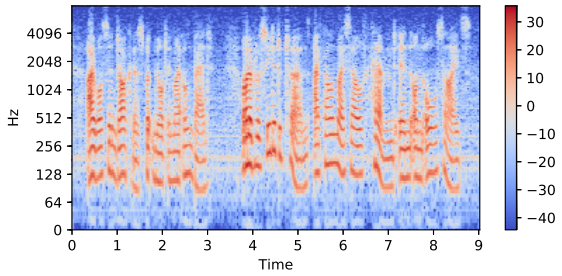
3.1.1. *WORLD Decoder*

The speech analysis and synthesis systems which provide high quality reconstruction after modification of parameters are useful in various applications like singing synthesis and voice conversion systems. In order to process large amounts of data, the algorithms need to be computationally efficient. The WORLD model is a fast algorithm which is vocoder based and generates high quality speech. The speech synthesis system consists of decomposing the speech signal into three components - fundamental frequency (F0) (which is estimated using the DIO algorithm [25]), spectral envelope (estimated using the CheapTrick algorithm [26]) and the excitation signal (estimated using the PLATINUM algorithm [27] denoted as aperiodic parameter). The F0 information is used to determine the temporal positions of the origin of the vocal cord vibrations. The F0 information in WORLD vocoder is estimated using a series of low-pass filters which gives multiple candidates. Using these mulitple candidate estimates for F0, a reliability measure based on the variance of the estimates is used to find the final F0 value.

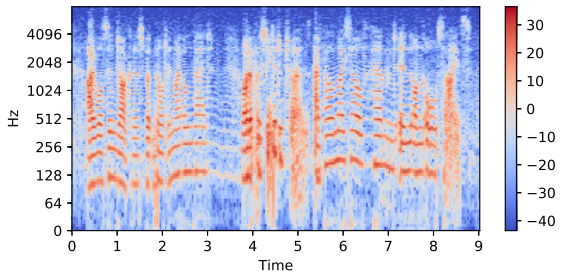
3.1.2. *V2S using WORLD*

The approach to data augmentation using the V2S model is shown in Figure 1. We use the WORLD vocoder to independently decompose the natural speech and singing voice (like opera vocals) into the constituent components. The F0 contour from the singing voice is then used along with the spectral envelope and the aperiodic parameter from the natural speech and fed to the synthesizer. The synthesized output is the singing voice version of the natural speech.

The western opera vocal dataset consists of both male and female opera singers and during the synthesis we make sure that the speech and the opera vocals are gender matched. Further, our analysis showed that, instead of randomly matching a natural speech recording with an opera vocal sample, a technique for choosing the opera vocal based on the closest average F0 value with that of the speech signal under consideration improved the quality of the synthesized output drastically. To facilitate this operation, we perform the decomposition of the opera vocals in the dataset apriori and also store the average F0 value. Then, for the given speech signal under consideration, the decomposition is performed and the average F0 value is computed. The opera vocal that has the closest average F0 value from the database is chosen and its F0 contour is used in the synthesis of the singing speech. We did not perform any alignment of the F0 track with the speech signal.



(a) Spectrogram of speech sample



(b) Spectrogram of synthesized waveform

Fig. 2: V2S singing voice conversion of natural speech.

An example illustration of the synthesized output from the proposed V2S approach is shown in Figure 2b. The original speech sample is also show here in Figure 2a. As seen here, the synthesized output has different harmonicity properties. However, phonemic activity of the speech is well preserved in the reconstructed output[1].

3.2. Transfer learning

Although V2S can provide large amounts of synthetic singing voice data and provides significant performance gain over the model trained without V2S data augmentation as shown in Section 5, there is still a domain mismatch between the real and synthetic singing voice since the spectral features from standard speech are used in V2S. However, large amount of monophonic singing voice data is not available, especially for singing voice of professional artists. To overcome this challenge, we propose a transfer learning approach using source-separated singing voice data. Since polyphonic music with transcriptions are relatively easy to obtain like the DALI corpus [17], we first separate a vocal track from polyphonic music using the state-of-the-art DNN-based source separation - D3Net [28]. Then, E2E lyric recognition model trained with V2S data augmentation is fine tuned on the source-separated (SS) singing voice data. One can also consider fine tuning the E2E models on polyphonic music data directly. However, we experimentally show in Section 5 that incorporating the singing voice source separation works significantly better for polyphonic and monophonic cases.

3.3. Language model for lyrics text

The probability distribution on words and their transition probability in lyrics are considerably different from standard speech. In particular, lyrics often contain artistic expressions and may potentially violate many grammar usage rules in natural speech. To incorporate the difference, we develop a language model using a combination of text from audio-books and lyrics data from DALI [17], Sony Singing Voice (SSV) dataset, and web resource[2]. This text data contained lyrics from a variety of genres like hip-hop, pop, classical, country-music, rock, jazz and consisted of about 5M lines of text.

[1]Some audio samples are available here - `https://github.com/iiscleap/V2S_Samples`

Table 1: Performance of different E2E architectures trained on natural speech from Librispeech corpus.

Architecture	SSV
Hybrid LSTM [13]	53.4
Transformer [14]	48.4

Table 2: WERs of models trained with and without V2S data augmentation. Here, SS-DALI denotes Source separated DALI, Poly-DALI denotes Polyphonic DALI. The results indicated as '–' correspond to conditions when some of the files fail to decode completely.

Models	SS-DALI		Poly-DALI		SSV
	Dev.	Test	Dev.	Test	
LS (original)	47.7	49.5	80.9	86.3	48.4
LS with V2S	43.7	46.8	–	–	41.6
LS with V2S (F0 map)	**42.8**	**44.9**	**75.2**	**78.9**	**39.7**

Table 3: Perplexity values on various test data for the LM. Here, LS refers to audio-books in the LibriSpeech test and dev set. The LM model in the last row is refered to as the Lyrics LM.

Training Data	LS	SSV	DALI
Audio-book	**28.7**	70.4	75.4
— fine-tuned on SSV+DALI+web	449.8	57.6	66.3
SSV+DALI+web	1002	75.9	81.6
Audio-book+SSV+DALI+web	73.0	**50.8**	**58.4**

4. EXPERIMENTAL SETUP

4.1. Dataset

We use the LibriSpeech corpus [29] as the natural speech data corpus in all our experiments. The LibriSpeech corpus contains 1000 hours of read speech sampled at 16 kHz. The training portion of the Librispeech contains 1129 female speakers and 1210 male speakers with each speaker providing 25-30 minutes of speech. The language model training data is also released as part of the Librispeech corpus which contains approximately $14,500$ public domain books with around 800M tokens in total and 900K unique words. We use the train, test and dev partitions as prescribed in the release [29].

For fine tuning to real singing voice, we use the DALI dataset [17]. This dataset consists of 1200 English polyphonic songs with lyrics-transcription, which results in a total duration of 70 hours. We split the dataset to a training, development, and test with a ratio of 8:1:1. The splits are also carefully performed not to have any overlap in the artists performing in each of these splits. Further, the splits are also gender balanced to avoid any bias in training/testing. For testing on monophonic singing voice to avoid the effect of accompaniment sounds, we also used a proprietary dataset from Sony Corp termed

[2]The lyrics data from www.lyrics.com

Table 4: WER values obtained after training the models on mapped modulated LibriSpeech and normal LibriSpeech. SS DALI - Source separated DALI, Poly DALI - Polyphonic DALI, SP - Speed Perturbation.

Model	Audiobook LM					Lyrics LM				
	SS-DALI		Poly-DALI		SSV	SS-DALI		Poly-DALI		SSV
	Dev.	Test	Dev.	Test		Dev.	Test	Dev.	Test	
No fine tuning	61.4	65.3	84.4	87.3	39.7	58.9	62.5	82.2	85.3	38.1
Fine tuned on SS-DALI	46.2	49.4	77.3	80.9	39.1	44.8	47.0	75.2	78.1	38.4
Fine tuned on SS-DALI (SP)	42.8	44.9	75.2	78.9	36.1	**41.5**	**43.4**	74.5	78.0	**34.8**
Fine tuned on Poly-DALI (SP)	56.0	58.7	59.1	61.8	45.5	55.1	57.4	**57.9**	**60.4**	44.8

as Sony Singing Voice (SSV) dataset which consists of 88 English Songs. Each of these songs have an approximate duration of 4 min and had a mix of multiple genres. We have used 5sec chunks of the audio for model training. The other model transformer model parameters follow the baseline Librispeech setup from ESPNET toolkit. The singing voice in these recordings is also of professional quality.

4.2. Language model

We train two different language models (LM) - an audio book LM using the text resources from the Librispeech corpus and a lyrics LM using a combination of audio-book text with 5M lines of lyrics text described in Sec. 3.3. The LMs are a recurrent neural network (RNN) model with 1 layer of 1024 LSTM cells. The language model is incorporated in the E2E system as described in [12].

4.3. E2E ASR Framework

We have used the ESPNET toolkit [30] to perform our E2E recognition experiments. The features used to train the architecture are log-filter bank features extracted using 30 mel-spaced windows of duration 25ms with a shift of 10ms. The E2E model used in most of the experiments is based on the Transformer architecture [14]. The encoder used is a 12-layer transformer network with 2048 units in the projection layer. The attention used is location attention and the decoder network is a 6-layer Transformer network with 2048 units in the projection layer. During training, multiple cost functions are used [14] like connectionist temporal cost (CTC) and the cross-entropy (CE) loss. The model is trained using Adam optimization and training is performed for several epoch till the loss saturates on the validation data. The CTC-weight is fixed at 0.3 and during decoding the beam-size is fixed at 20. Both the LMs (audio-book LM as well as the lyrics LM) have the same architecture and used 5000 sub-word units at the output layer.

5. RESULTS

E2E model architecture: The first set of results shown in Table 1 highlights the lyrics transcription results on the SSV data using the Hybrid LSTM E2E architecture [13] and the transformer architecture [14]. Both the models are trained on natural speech. As seen in this Table, the transformer architecture provides improved robustness to singing voice transcription task even when the model is trained for speech recognition task [14].

V2S: Table 2 shows the impact of the proposed V2S data augmentation for training the E2E model. The E2E model trained on natural speech provides very high word-error-rates (WER) on polyphonic DALI dataset. The V2S improves the lyrics transcription performance on the SSV data significantly (average relative improvements of 14% over the natural speech based WER). Further, the V2S applied with F0 mapping, where the opera vocals are selected to match the average F0 of the speech file under consideration, further improves the lyrics transcription accuracy. These results validate the effectiveness of proposed V2S data augmentation and F0 mapping (relative improvements of 18 % on the SSV dataset).

LM comparison: The comparison of various language models in terms of perplexity values is shown in Table 3. The audio-book LM refers to the LM trained using the text data from the Librispeech corpus. The perplexity values highlighted here suggest that, training the LM from the mixed corpus of speech and lyrics (from SSV, DALI and web) provides the best perplexity compared to either fine-tuning the audio-book LM or using only lyrics text for LM training. The best model (last row) in Table 3 is referred to as the lyrics LM in the experiments that follow. Further, the fine-tuned version of the audio-book LM (second row) is referred to as the audio-book LM for the remainder of the experiments.

Transfer learning: The impact of fine-tuning the E2E models which were trained with V2S is shown in Table 4. These results show that fine-tuning on the source separated (SS) DALI training data can improve the performance on the DALI test data. Further, the application of speed perturbation in training [31] improves the lyrics transcription performance on both DALI and SSV test data. The application of V2S data augmentation in model training along with speed perturbation improves the lyrics transcription WER of SSV data relatively by about 28 % over the natural speech based E2E system. In addition, the WER results for the final system of 34.8 % WER on SSV data and 43.4 % on DALI data suggest that the highly challenging task of lyrics transcription on monophonic/polyphonic music with large variety of music genres can be explored with partial success using the techniques proposed in this work.

6. SUMMARY

This paper presents a data augmentation method for E2E lyric transcription. The proposed method, termed as voice-to-singing (V2S), modulates the natural speech to a singing style by replacing a fundamental frequency contour of natural speech to that of singing voices using a vocoder based speech synthesizer. We also propose a transfer learning based approach to leverage a large amount of source-separated real singing voices from polyphonic music. The application of proposed methods are explored in the design of lyrics transcription system based on transformer based E2E model. Various experiments highlight the performance benefits of using the proposed V2S along with lyrics language modeling and transfer learning. E2E ASR model is shown to provide useful features for other applications such as singing voice separation [32]. Adopting the E2E lyric recognition model to such music applications is one of our future works.

7. REFERENCES

[1] S. O. Ali and Z. F. Peynircioğlu, "Songs and emotions: are lyrics and melodies equal partners?," *Psychology of music*, vol. 34, no. 4, pp. 511–534, 2006.

[2] A. J. Good, F. A. Russo, and J. Sullivan, "The efficacy of singing in foreign-language learning," *Psychology of Music*, vol. 43, no. 5, pp. 627–640, 2015.

[3] C. Gupta, E. Yılmaz, and H. Li, "Acoustic modeling for automatic lyrics-to-audio alignment," *arXiv preprint arXiv:1906.10369*, 2019.

[4] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–11, 2010.

[5] G. Dzhambazov et al., *Knowledge-based probabilistic modeling for tracking lyrics in music audio signals*, Ph.D. thesis, Universitat Pompeu Fabra, 2017.

[6] T. Hosoya, M. Suzuki, A. Ito, S. Makino, L. A. Smith, D. Bainbridge, and I. H. Witten, "Lyrics recognition from a singing voice based on finite state automaton for music information retrieval.," in *ISMIR*, 2005, pp. 532–535.

[7] H. Fujihara, M. Goto, and J. Ogata, "Hyperlinking lyrics: A method for creating hyperlinks between phrases in song lyrics.," in *ISMIR*, 2008, pp. 281–286.

[8] M. Gruhne, C. Dittmar, and K. Schmidt, "Phoneme recognition in popular music.," in *ISMIR*, 2007, pp. 369–370.

[9] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1252–1261, 2011.

[10] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[11] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764–1772.

[12] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *IEEE ICASSP*. IEEE, 2016, pp. 4945–4949.

[13] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[14] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, et al., "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 449–456.

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[16] M. MORISE, F. YOKOMORI, and K. OZAWA, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[17] G. Meseguer Brocal, "The dali dataset," Feb. 2019.

[18] C. Gupta, R. Tong, H. Li, and Y. Wang, "Semi-supervised lyrics and solo-singing alignment," in *ISMIR*, 2018.

[19] B. Sharma, C. Gupta, H. Li, and Y. Wang, "Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models," in *IEEE ICASSP*, 2019, pp. 396–400.

[20] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *IEEE ICASSP*, 2019, pp. 181–185.

[21] C. Gupta, E. Yilmaz, and H. Li, "Acoustic modeling for automatic lyrics-to-audio alignment," in *INTERSPEECH*, 2019.

[22] C. Gupta, E. Yılmaz, and H. Li, "Automatic lyrics alignment and transcription in polyphonic music: Does background music help?," in *IEEE ICASSP*, 2020, pp. 496–500.

[23] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *IEEE WASPAA*. IEEE, 2007, pp. 215–218.

[24] J. Parekh, P. Rao, and Y.-H. Yang, "Speech-to-singing conversion in an encoder-decoder framework," in *IEEE ICASSP*. IEEE, 2020, pp. 261–265.

[25] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009.

[26] M. Morise, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1 – 7, 2015.

[27] M. Morise, "Platinum: A method to extract excitation signals for voice synthesis system," *Acoustical Science and Technology*, vol. 33, no. 2, pp. 123–125, 2012.

[28] N. Takahashi and yuki Mitsufuji, "D3net: Densely connected multidilated densenet for music source separation," *CoRR, preprint https://arxiv.org/abs/2010.01733*, 2020.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE ICASSP*, 2015, pp. 5206–5210.

[30] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," 2018.

[31] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[32] N. Takahashi, M. K. Singh, S. Basak, P. Sudarsanam, S. Ganapathy, and Y. Mitsufuji, "Improving Voice Separation by Incorporating End-To-End Speech Recognition," in *Proc. ICASSP*, 2020.